Trouver et confondre les coupables : un processus sophistiqué de correction de lexique

Lionel Nicolas, Benoît Sagot, Miguel A. Molinero, Jacques Farré, Éric Villemonte de La Clergerie

Trouver et confondre les coupables : un processus sophistiqué de correction de lexique *

Lionel Nicolas⋆, Benoît Sagot⊕, Miguel A. Molinero◇,
Jacques Farré⋆, Éric de La Clergerie⊕.

⋆ Équipe RL, Laboratoire I3S, UNSA+CNRS, France {lnicolas,jf}@i3s.unice.fr

◇ Grupo LYS, Univ. de A Coruña, España mmolinero@udc.es

⊕ Projet ALPAGE, INRIA Rocquencourt + Paris 7, France {benoit.sagot, Eric.De_La_Clergerie}@inria.fr

Résumé. La couverture d'un analyseur syntaxique dépend avant tout de la grammaire et du lexique sur lequel il repose. Le développement d'un lexique complet et précis est une tâche ardue et de longue haleine, surtout lorsque le lexique atteint un certain niveau de qualité et de couverture. Dans cet article, nous présentons un processus capable de détecter automatiquement les entrées manquantes ou incomplètes d'un lexique, et de suggérer des corrections pour ces entrées. La détection se réalise au moyen de deux techniques reposant soit sur un modèle statistique, soit sur les informations fournies par un étiqueteur syntaxique. Les hypothèses de corrections pour les entrées lexicales détectées sont générées en étudiant les modifications qui permettent d'améliorer le taux d'analyse des phrases dans lesquelles ces entrées apparaissent. Le processus global met en oeuvre plusieurs techniques utilisant divers outils tels que des étiqueteurs et des analyseurs syntaxiques ou des classifieurs d'entropie. Son application au Le*fff*, un lexique morphologique et syntaxique à large couverture du français, nous a déjà permis de réaliser des améliorations notables.

Abstract. The coverage of a parser depends mostly on the quality of the underlying grammar and lexicon. The development of a lexicon both complete and accurate is an intricate and demanding task, overall when achieving a certain level of quality and coverage. We introduce an automatic process able to detect missing or incomplete entries in a lexicon, and to suggest corrections hypotheses for these entries. The detection of dubious lexical entries is tackled by two techniques relying either on a specific statistical model, or on the information provided by a part-of-speech tagger. The generation of correction hypotheses for the detected entries is achieved by studying which modifications could improve the parse rate of the sentences in which the entries occur. This process brings together various techniques based on different tools such as taggers, parsers and entropy classifiers. Applying it on the Le*fff*, a large-coverage morphological and syntactic French lexicon, has already allowed us to perfom noticeable improvements.

Mots-clés : Acquisition et correction lexicale, lexique à large couverture, fouille d'erreurs, étiqueteur syntaxique, classifieur d'entropie, analyseur syntaxique.

Keywords: Lexical acquisition and correction, wide coverage lexicon, error mining, tagger, entropy classifier, syntactic parser.

*Ces travaux ont notamment pu être réalisés grâce au soutient du ministère de l'éducation et des sciences d'Espagne, FEDER (HUM2007-66607-C04-02), du Gouvernement Régional de Galice (INCITE08PXIB302179PR, INCITE08E1R104022ES) et du *Galician Network for Language Processing and Information Retrieval* 2006-2009.

1 Introduction

Le développement manuel d'un lexique précis et à large couverture est une tâche fastidieuse, complexe et sujette à erreurs nécessitant une coûteuse expertise humaine. Les développements manuels de lexiques n'atteignent généralement pas les objectifs attendus et progressent très lentement une fois un certain niveau de couverture et de qualité atteint. Cette tâche manuelle peut cependant être simplifiée et améliorée par l'utilisation d'outils automatisant les tâches d'acquisition et de correction. Nous présentons un ensemble combiné de techniques permettant de détecter les entrées manquantes, incomplètes ou erronées d'un lexique et de proposer des corrections. La chaîne logique du processus global se résume ainsi :

1. Donner en entrée à un analyseur syntaxique un grand nombre de phrases non-annotées considérées comme respectueuses de la langue, afin d'attribuer un échec d'analyse aux manques de l'analyseur et non aux textes qu'il reçoit en entrée[1].
2. Pour chaque phrase non-analysable, tenter de déterminer automatiquement si l'échec d'analyse est dû à des manques de la grammaire ou du lexique utilisés par l'analyseur.
3. Suspecter des entrées lexicales d'être manquantes, incomplètes ou erronées.
4. Générer des hypothèses de correction en observant les attentes de la grammaire vis à vis des formes suspectées lors des analyses de phrases dans lesquelles elles apparaissent.
5. Évaluer et classer les hypothèses de correction afin de procéder à une validation manuelle.

Bien que tous nos exemples et résultats soient liés à la langue française, cet ensemble de techniques est indépendant du système, c.a.d, il est facilement adaptable à la plupart des étiqueteurs syntaxiques, classifieurs d'entropie, lexiques et analyseurs profonds existants, et par conséquent, à la plupart des langues informatiquement décrites.

Cet ensemble de techniques est l'un des points de départ du récent projet Victoria[2], dont le but est de développer un ensemble d'outils permettant la construction efficiente de ressources morphologiques, lexicales et grammaticales. Ces travaux font suite aux travaux présentés dans (2007a; 2007b; 2008) où des modèles plus simples avait été décrits.

Pour des raisons de clarté, les résultats pratiques de chaque étape ainsi que les améliorations possibles sont donnés conjointement à sa présentation. Nous commençons donc par décrire le contexte pratique de nos expériences (sect. 2). Nous décrivons ensuite chaque étape énumérée ci dessus (sections 3,4,5,6,7) et concluons (sect.8).

2 Contexte pratique

Nous utilisons un corpus journalistique français non-annoté extrait du *monde diplomatique*. Ce corpus contient 280 000 phrases de 25 mots ou moins, totalisant 4,3 millions de mots.

Le lexique utilisé et amélioré se nomme le Le*fff*[3]. Ce lexique morphologique et syntaxique à large couverture du français contenant plus de 600 000 entrées.

Deux analyseurs syntaxiques sont utilisés afin de générer des corrections :

[1] textes de lois, journaux, etc.

[2] `http://www.victoria-project.org`, octobre 2008.

[3] Lexique des formes fléchies du français. `http://alpage.inria.fr/~sagot/lefff-en.html`.

– FRMG (***Fr**ench **M**eta-**G**rammar*) se base sur une méta-grammaire abstraite avec des arbres hautement factorisés (Thomasset & Villemonte de La Clergerie, 2005) compilée en un analyseur hybride TAG/TIG grâce au système DYALOG.
– SXLFG-FR (Boullier & Sagot, 2006) est une grammaire LFG profonde efficace non probabiliste compilée en analyseur LFG par SXLFG, un système basé sur SYNTAX.

Nous utilisons aussi de façon ponctuelle l'étiqueteur syntaxique MrTagoo (Molinero *et al.*, 2007; Graña, 2000) et le classifieur d'entropie MegaM (Daumé III, 2004).

3 Classification des phrases non analysables

Nous partons des résultats d'analyse syntaxique d'un grand nombre de phrases. Certaines phrases ont été pu être analysées, d'autres non. Les phrases analysables sont considérées comme couvertes lexicalement et grammaticalement (même si les analyses obtenues ne coïncident pas toujours avec leur sens véritable). Les phrases non-analysables sont par contre non couvertes lexicalement et/ou grammaticalement. Afin de générer des corrections pour un lexique, il nous est préférable d'isoler les phrases qui ne sont non-analysables que pour des raisons lexicales. Pour ce faire, nous cherchons d'abord à identifier les phrases grammaticalement non-analysables.

Cette étape est réalisée grâce à un classifieur d'entropie, c.a.d, un outil statistique qui permet de calculer une adéquation (une entropie) entre les données qu'il reçoit à l'évaluation et les données sur lesquelles il a été entraîné (Daumé III, 2004).

Dans ce but, nous profitons du fait que les constructions syntaxiques sont plus fréquentes et bien moins diverses que les formes lexicales. Celles non couvertes tendent donc à être récurrentes et systématiques dans les phrases grammaticalement non-analysables. Afin d'identifier ces constructions problématiques, nous entraînons un classifieur d'entropie de la sorte :
– nous réduisons toutes les phrases à des séquences de 3-grams obtenues soit à partir des catégories syntaxiques pour les mots des catégories ouvertes (c.a.d., verbes, adjectifs, etc.) et soit à partir de formes lexicales pour les mots des catégories fermées (prépositions, déterminants, etc.) auxquelles nous rajoutons des marqueurs de début et de fin de phrase.
– nous associons à chaque séquence une classe (*analysable/non-analysable*) correspondant au résultat de l'analyse de la phrase dont cette séquence a été extraite.

Pour "je(cln) mange(v) une(det) pomme(nc)" nous générons donc une séquence de 3-grams d'entraînement: <deb-je-v> <je-v-une> <v-une-nc> <une-nc-fin> dont la classe est *analysable*.

Le classifieur différencie donc, à partir des 3-grams qui les composent, les phrases qui paraissent être grammaticalement analysables de celles qui ne le sont pas. Les phrases non-analysables déclarées comme grammaticalement analysables sont alors considérées comme *lexicalement non-analysables*.

Il est à noter que l'entraînement n'est pas optimal à cause de deux aspects. Premièrement, la catégorie de chaque mot dans les phrases est obtenue par le biais d'un étiqueteur syntaxique. Les étiqueteurs ne sont clairement pas des outils parfait. Cependant, leurs erreurs sont grammaticalement aléatoires car elles dépendent avant tout des formes lexicales rencontrées. Ce caractère aléatoire permet donc aux erreurs de ne pas trop perturber la cohérence globale de l'entraînement du classifieur d'entropie. Deuxièmement, les phrases non-analysables données en entraînement ne sont pas toutes grammaticalement non-analysables, certaines sont seulement lexicalement non-analysables. On l'entraîne donc en partie à considérer injustement des phrases comme grammaticalement non-analysables. Cependant, les calculs sur les 3-grams pré-

sents dans ces phrases injustement catégorisées sont contrebalancés par leur présence logique dans des phrases analysables.

Pour évaluer cette technique, nous avons ôté 5% des phrases analysables à l'entraînement et avons observé si le classifieur les déclare comme analysables. Les taux de précision avant la première session de correction, puis après la première, seconde et troisième session étaient respectivement de 92,7%, 93,8%, 94,1% et 94,9%. La précision du classifieur augmente logiquement car, après chaque session, certaines phrases dont l'analyse échouait pour des raisons lexicales deviennent analysables et ne perturbent donc plus l'entraînement. La génération des séquences de 3-grams étant la même pour l'ensemble des phrases, ces taux de précision devraient s'appliquer de façon équivalente aux phrases grammaticalement non-analysables.

Finalement, le taux d'erreur de (pour l'instant) 5,1% est un manque considéré comme acceptable étant donné l'impact positif que l'étape de filtrage a sur nos techniques de détection. Puisqu'il n'y a pas de raison pour qu'une forme particulière se retrouve plus que de raison dans des phrases classifiées incorrectement, il est possible de contrebalancer la perte de certaines phrases par une simple augmentation de la taille du corpus donné en entrée.

4 Détection des manques lexicaux

La détection d'entrées lexicales douteuses est réalisée par le bais de deux techniques complémentaires qui identifient des formes et les associent à des phrases dont elles sont suspectées d'être responsables de l'échec d'analyse.

4.1 Détection d'information lexicale à courte portée via un étiqueteur

Nous appelons information lexicale de courte portée toute information pouvant être déterminée par un étiqueteur syntaxique. Pour l'instant, nous ne considérons que la catégorie syntaxique.

Afin de détecter les problèmes lexicaux concernant ce type d'information, nous utilisons un étiqueteur syntaxique configuré de façon particulière dont nous court-circuitons ponctuellement le lexique interne afin de le forcer à considérer comme inconnue, une à la fois, chaque forme d'une phrase. Nous nous reposons donc sur sa capacité à s'inspirer du contexte d'une forme pour supposer l'étiquette la plus probable. Les informations portées par ces étiquettes supposées sont ensuite comparées aux informations existantes dans le lexique. Si ces informations sont manquantes et concernent des classes ouvertes, la forme correspondante est déclarée comme suspecte. Appliquée aux catégories syntaxiques, cette technique nous permet de détecter les homonymes manquants d'un lexique en plus des formes totalement inconnues.

Bien entendu, les étiqueteurs commettent des erreurs, surtout lorsqu'on courcuite ainsi leurs lexiques internes. La précision de l'étiqueteur modifié pour n'importe quel type de forme (même celles appartenant aux classes fermées) sur 5% des phrases non utilisées à l'entraînement est de 47,34%. Le nombre de faux positifs est donc très important. En étudiant les résultats, nous avons pu observer le caractère systématique de certaines erreurs. Par exemple, un nom propre est souvent considéré comme un nom commun, un participe passé comme un adjectif etc. Pour réduire le nombres de formes suspectées incorrectement, nous avons développé quatre surcouches profitant du fait que nous ne travaillions pas à l'échelle d'une phrase solitaire (comme le font généralement les étiqueteurs) mais d'un ensemble de phrases.

La première surcouche choisie simplement l'étiquette la plus fréquement donnée à une forme.

La deuxième surcouche calcule des patrons de réponses de l'étiqueteur par type d'étiquette et par fréquence d'apparition de la forme (indexées sur les valeurs entières du logarithme népérien). On cherche donc durant l'entraînement à savoir, par exemple, combien de fois un nom propre est déclaré par l'étiqueteur comme nom propre, comme nom commun, comme adjectif, etc. A l'évaluation, nous calculons une affinité entre les nouveaux ensembles de réponses données par l'étiqueteur pour chaque forme et les patrons calculés à l'entraînement et on attribue ensuite à la forme l'étiquette du patron avec lequel elle a le plus d'affinité. Le calcul d'affinité $Aff_{pat/rep}$ et le choix du meilleur patron $Best_{pat}$ se calculent ainsi :

$$Aff_{pat/rep} = \sum abs(Pat_{eti} - Rep_{eti}), Best_{pat} = max(Aff_{pat/rep} * log(Occ_{pat}))$$

Pat_{eti} et Rep_{eti} sont la part d'une étiquette eti, et Occ_{pat} est le poids d'émission du patron égale à la somme des occurrences des formes qui ont permis sa construction.

La troisième surcouche applique la même idée mais laisse le calcul d'affinité à un classifieur d'entropie. Le classifieur est entraîné à reconnaître des patrons et à les associer à une classe représentant une étiquette et un index népérien.

La dernière surcouche s'appuie sur les trois premières pour réaliser un « vote à la crédibilité »où l'« opinion » de chaque surcouche est valorisée à partir des taux d'erreurs par type de réponse.

Le défaut majeur de ces surcouches est qu'actuellement, elles considèrent que chaque forme représente un seule lemme, ce qui est faux bien que vrai dans la grande majorité des cas.

L'étiqueteur est alors entraîné avec 50% des phrases du corpus, l'entraînement des surcouches se réalise ensuite sur les réponses fournis par l'évaluation de l'étiqueteur sur 47,5% des phrases et leur évaluation sur 2,5% des phrases restantes. Les précisions respectives sur ces 2,5% de phrases de l'étiqueteur, de la première, seconde, troisième et quatrième surcouche sont respectivement de 40,17%, 43,6%, 77,61%, 74,09%[4] et 89,78%. L'application de ces surcouches permet donc de passer d'une précision originelle de 47,34% de l'étiqueteur modifié à 89,78%, réduisant ainsi fortement le nombre de faux positifs.

Dans une première version basée uniquement sur l'étiqueteur sans surcouche et considérant toute les formes comme inconnues en même temps, nous avons pu identifier 182 lemmes manquants. Cette nouvelle version nous a permis d'en trouver 358 autres. Le tout correspond a un total de 1168 formes lexicales, pour la plupart adjectifs ou noms propres manquants.

4.2 Approche statistique pour la détection de défauts lexicaux

Cette technique de détection de défauts lexicaux, décrite dans (Sagot & Villemonte de La Clergerie, 2006; Sagot & de La Clergerie, 2008), repose sur les hypothèses suivantes :

– Si une forme lexicale apparaît plus souvent dans des phrases non-analysables que dans des phrases analysables, il est raisonnable de la suspecter d'être incorrectement décrite dans le lexique (van Noord, 2004).
– Le taux de suspicion peut être renforcé si la forme apparaît dans des phrases non-analysables à côté d'autres formes présentes dans des phrases analysables.

[4]Ce résultat moins important que la précédente surcouche est probablement dû à une configuration insuffisante du classifieur d'entropie

L'avantage de cette technique par rapport à la précédente est sa capacité à prendre en compte tout type d'erreurs lexicales. Cependant, puisque qu'elle part du précepte que toute phrase non-analysable ne l'est que pour des raisons lexicales, la qualité de la liste de suspects fournie dépend directement de la qualité de la grammaire utilisée. En effet, si une forme spécifique est particulièrement liée à une construction syntaxique non couverte par la grammaire, on la retrouvera souvent dans des phrases non analysables et elle sera alors injustement suspectée.

Nous atténuons ce problème de deux façons. Premièrement, nous excluons du calcul statistique toutes les phrases considérées comme grammaticalement non-analysables. Deuxièmement, comme cela a déjà été fait dans (Sagot & Villemonte de La Clergerie, 2006), nous combinons les résultats d'analyse fournis par différents analyseurs reposant sur des formalismes et grammaires différents, et donc avec des manques grammaticaux différents.

Cette technique nous a permis de détecter 72 lemmes décrits de façon incomplète correspondant à un total de 1693 formes lexicales, pour la plupart des verbes.

Pour l'instant, les deux techniques de détection identifient des formes lexicales. Il serait intéressant de monter à niveau de lemme en appliquant en post-traitement l'idée décrite dans (Sagot, 2005) où la validité d'un lemme est favorisée ou pénalisée suivant la présence ou l'absence de ses formes lexicales.

Autre amélioration possible, l'efficacité/l'intérêt de ces techniques ne sont valorisés que part les corrections qu'elles ont permis de réaliser. Il serait intéressant d'établir une métrique afin d'évaluer la qualité des formes suspectes fournies. Cette métrique permettrait aussi de quantifier formellement l'impact positif de l'étape de filtrage sur l'étape de détection. Le classement à chaque session des formes corrigées pourrait être un point de départ.

5 Génération des hypothèses de correction lexicale : analyse de phrases initialement non-analysables

La génération de corrections lexicales à partir du contexte grammatical a été utilisé pour la première fois en 1990 (Erbach, 1990). Elle suit l'idée suivante : suivant la qualité du lexique et de la grammaire, la probabilité que ces deux ressources soient simultanément erronées au sujet d'une forme donnée dans une phrase donnée peut être faible. Si une phrase ne peut pas être analysée à cause d'une forme suspecte, cela implique que les deux ressources n'ont pas pu s'accorder sur le rôle que la forme peut avoir dans la phrase. Puisque que le problème est d'origine lexical, il est possible de générer des corrections en étudiant les attentes de la grammaire pour chaque forme suspectée lorsqu'elle analyse les phrases qui leur sont associées. De manière métaphorique, on « demande » à la grammaire son opinion sur les formes suspectées. Originellement, les formes suspectes étaient déterminées manuellement puis, à partir de 2006 (van de Cruys, 2006; Yi & Kordoni, 2006; Nicolas *et al.*, 2007a; Nicolas *et al.*, 2007b) , cette idée a été combinée avec des techniques de fouille d'erreurs telles que (van Noord, 2004; Sagot & Villemonte de La Clergerie, 2006; Sagot & de La Clergerie, 2008).

Pour générer des corrections, nous nous approchons au mieux de des analyses que la grammaire aurait permises avec un lexique sans erreur. Puisque nous pensons que les informations lexicales associées à la forme suspecte ont coupé le chemin vers une possible analyse, nous diminuons les restrictions imposées par les informations lexicales : pendant l'analyse, chaque fois qu'une information lexicale associée à une forme suspectée est vérifiée, le lexique est court-

circuité et toutes les contraintes sont considérées comme satisfaites. La forme devient alors tout ce que peut souhaiter la grammaire. En réalité, cette opération est effectuée en échangeant les formes suspectes dans les phrases associées par des formes sous-spécifiées appelées *jokers*. Si une forme a été correctement suspectée, et si c'est l'unique cause d'échec de certaines analyses de phrases, remplacer cette forme par un joker permet aux phrases de devenir analysables. Dans ces nouvelles analyses, des entrées « instanciées » du joker sont partie prenante des structures grammaticales produites en sortie. Ces entrées instanciées représentent les informations manquantes du lexique, nous les traduisons donc au format du lexique afin d'établir les corrections lexicales.

Comme expliqué dans (Barg & Walther, 1998), l'utilisation de jokers totalement sous-spécifiés peut introduire une ambiguïté trop grande dans le processus d'analyse. Cela entraîne souvent des échecs d'analyse pour des contraintes de temps ou de mémoire (pas de corrections), ou des analyses surgénérative (trop de corrections). Contrairement à (Yi & Kordoni, 2006), où les auteurs utilisent les jokers totalement spécifiés les plus probables, nous n'ajoutons que peu d'information lexicale aux jokers et nous nous reposons sur la capacité de nos analyseurs à gérer des formes sous spécifiées. Pour des raisons pratiques, nous avons choisi d'ajouter aux jokers une catégorie syntaxique. L'ambiguïté introduite reste conséquente et aboutit généralement à un nombre important de corrections. Néanmoins cet aspect peut être facilement contrebalancé pour peu qu'il y ait assez de phrases non-analysables associées à une forme suspecte (voir sect 6). La catégorie syntaxique ajoutée aux jokers dépend de la technique de détection utilisée pour suspecter la forme. Lorsque nous utilisons la détection basée sur un étiqueteur, nous générons des jokers avec des catégories syntaxiques en accord avec les étiquettes fournies pour la forme. Quand nous utilisons l'approche de détection statistique, nous produisons des jokers avec les catégories syntaxiques déjà présentes dans le lexique pour la forme suspectée.

6 Extraction et classement des corrections

Le lecteur a pu noter qu'un joker inadéquat peut parfaitement mener à de nouvelles analyses et donc permettre la génération de corrections incorrectes. Nous séparons donc les corrections suivant le joker qui a permis leur génération.

puis les classons en accord avec les idées suivantes.

Classification mono-analyseur. À l'échelle d'une seule phrase, rien de permet de différencier les corrections valides des corrections erronées dont la génération résulte de l'ambiguïté introduite par les jokers. Cette ambiguïté ayant permis à l'analyse d'emprunter des règles de grammaires qu'elle n'aurait pas du. Le type de correction erronée dépend donc des règles de grammaires empruntées, c.a.d, de la structure syntaxique véritable de la phrase. Si la forme corrigée appartient à une catégorie ouverte, elle a de forte chance de pouvoir se retrouver au sein de structures variées. Par conséquent, plus le nombre de phrases est élevé, plus les structures syntaxiques au sein desquelles la forme est présente sont variées et plus les corrections erronées ont tendance a se disperser. Les correction valides, au contraire, tendent à être récurrentes.

Nous considérons donc toutes les corrections d'une forme w issue d'une même phrase comme un *groupe* de corrections. Chaque groupe reçoit un poids $P = c^n$ variant selon sa taille n, avec c une constante numérique entre $]0, 1[$ proche de 1. Plus le groupe est grand, plus bas sera

son poids car plus forte sera la probabilité qu'il soit la conséquence de squelettes syntaxiques *permissifs*. Chaque correction σ du groupe reçoit ensuite un poids $p_{g\sigma} = \frac{P}{n} = \frac{c^n}{n}$. Tous les poids d'une correction sont finalement additionnés afin de calculer le poids global $s_\sigma = \Sigma_g p_{g\sigma}$.

Classification multi-analyseur. Étant donné que les corrections erronées générées dépendent des règles de grammaire empruntées durant les analyses, l'utilisation des résultats provenant de plusieurs analyseurs avec des grammaires différentes permet d'accentuer leur dispersion, alors que les correction pertinentes restent habituellement stables. Des corrections sont donc considérées comme moins pertinentes si elles ne sont pas proposées par l'ensemble des analyseurs. Nous obtenons donc séparément les corrections de chaque analyseur comme décrit ci dessus et fusionnons les résultats à l'aide d'une simple moyenne harmonique.

7 Validation manuelle des corrections

Contrairement à (van de Cruys, 2006; Yi & Kordoni, 2006), nous privilégions une approche semi-automatique impliquant une étape de validation manuelle. Lors de la validation manuelle, nous avons identifié trois situations possibles.

Soit il n'y a pas de corrections : la détection des formes suspectes a été inadéquate ou la forme suspectée n'est pas l'unique raison des échecs d'analyse associés.

Soit il y a des corrections pertinentes : la forme a été correctement détectée, la forme est l'unique raison de (certains) échecs d'analyse associés.

Soit il n'y a que des corrections erronées : l'ambiguïté introduite par les jokers a ouvert la voie vers des analyses erronées fournissant des corrections erronées. Si la grammaire ne couvre pas toutes les structures syntaxiques possibles, il n'y a aucune garantie qu'il y ait des corrections pertinentes produites.

Les résultats donnés dans (van de Cruys, 2006) démontrent clairement cet aspect : on peut y voir que pour les catégories syntaxiques complexes comme les verbes, il est impossible d'appliquer un tel ensemble de techniques de façon automatisée sans nuire à la qualité du lexique. Si le but du processus de correction est d'améliorer la qualité du lexique et non pas d'augmenter artificiellement sa couverture, un tel processus devrait toujours être semi-automatique.

Comme nous l'expliquons plus loin, la validation manuelle n'est pas un très lourd tribut à payer. De plus, elle ouvre la possibilité suivante : les lemmes sémantiquement reliés d'une même catégorie syntaxique tendent à avoir des comportements syntaxiques similaires. Cette similarité pourrait être utilisée pour attirer l'attention du correcteur ou même générer des corrections pour des formes non rencontrés/détectés.

Voici quelques exemples de correction validées :

- *italien, portugais, parabolique, pittoresque, minutieux* étaient des adjectifs manquants ;
- *revenir* ne traitait pas les constructions telles que *revenir vers* ou *revenir de* ;
- *partager* ne traitait pas les constructions telles que *partager (quelque chose) entre* ;
- *ériger* était décrit comme attendant obligatoirement un COD et un attribut ;
- *livrer* ne traitait pas les constructions telles que *livrer (quelque chose) à quelqu'un*.

Le tableau 1 donne les résultats de 4 sessions de correction. Les première et troisième sessions ont été réalisées avec la technique statistique de détection. La seconde avec une version brute de la technique de détection basée sur un étiqueteur et la quatrième avec la version décrite précédemment (voir sect 4).

Après ces quelques sessions, les techniques de détections nous fournissent encore des formes suspectes mais nous n'obtenons plus de nouvelles corrections valides. Cela peut s'expliquer par

Session	1	2	3	4	total
nc	30	99	1	6	136
adj	66	694	27	14	801
verbs	1183	0	385	0	1568
adv	1	7	0	0	8
np	0	0	0	348	348
total	1280	800	413	368	2861

TAB. 1 – Formes lexicales mises à jour à chaque session.

plusieurs raisons. Bien que peu probable, les phrases non analysables restantes peuvent posséder deux formes erronées ; l'introduction d'un seul joker ne suffit donc pas à rendre la phrase analysable. On peut aussi penser que les couvertures de nos grammaires sont insuffisantes, elles ne sont donc pas en mesure de nous fournir de nouvelles corrections. Cette dernière explication est privilégiée car, après la dernière session, l'étape de filtrage des phrases non analysables a classifiée l'essentiel des phrases restantes comme grammaticalement non analysables. Des sessions de correction futures n'auront donc de sens qu'après des améliorations des grammaires ou l'application à de nouveaux corpus.

Cependant, cette constatation nous met en mesure de produire des corpus globalement représentatifs de manques grammaticaux. Si une technique était capable d'utiliser ce corpus pour suggérer des corrections grammaticales, la mise à jour de la grammaire nous permettrait de générer à nouveau des corrections pour le lexique. Ce qui à nouveau nous permettrait de générer un corpus représentatif des manques de la grammaire et ainsi de suite. Il serait alors possible de mettre au point un processus itératif améliorant alternativement et incrémentalement la grammaire et le lexique. Le modèle d'entropie construit par le classificateur pourrait être un bon point de départ pour établir les manques d'une grammaire.

Pour résumer nos résultats, nous avons déjà détecté et corrigé 612 lemmes correspondant à 2861 formes. Il est important de noter que ces corrections ont été obtenues après seulement quelques heures de travail manuel. L'aspect semi-automatique de notre approche n'est donc pas un très lourd tribut à payer.

8 Conclusion

Depuis ses premières versions (Nicolas *et al.*, 2007a; Nicolas *et al.*, 2007b), cet ensemble de techniques a fortement évolué et les résultats obtenus démontrent sa cohérence et sa viabilité. Les améliorations prévues devraient renforcer ces résultats et accroître l'efficacité globale. Une effort important sera de formaliser l'intérêt et l'efficacité des techniques par des métriques qui, à ce jour, n'existent pas pour ce type de problème.

Pour conclure, cet ensemble de techniques présente actuellement trois avantages importants :

1. Il prend en entrée du texte « non-annoté » produits quotidiennement par des sources journalistiques ou techniques facilement accessibles à travers des initiatives tel que le projet français Passage[5], qui juxtapose des fragments du Wikipedia français, de sources Wiki français, du journal régional *L'Est Républicain*, d'Europarl et de JRC Acquis.

[5] `http://atoll.inria.fr/passage`.

2. Il permet d'améliorer de façon significative un lexique morphologique et syntaxique à large couverture en peu de temps.
3. Enfin, son application répétée sur un corpus peut rendre ce corpus représentatif des manques de la grammaire utilisée. Un tel corpus pourrait être un point de départ pour le développement d'un processus dédié à l'amélioration d'une grammaire.

Références

BARG P. & WALTHER M. (1998). Processing unkonwn words in hpsg. In *Proceedings of the 36th Conference of the ACL and the 17th International Conference on Computational Linguistics*.

BOULLIER P. & SAGOT B. (2006). Efficient parsing of large corpora with a deep LFG parser. In *Proceedings of LREC'06*.

DAUMÉ III H. (2004). Notes on CG and LM-BFGS optimization of logistic regression. Paper http://pub.hal3.name/daume04cg-bfgs, implementation http://hal3.name/megam/.

ERBACH G. (1990). Syntactic processing of unknown words. In *IWBS Report 131*.

GRAÑA J. (2000). *Técnicas de Análisis Sintáctico Robusto para la Etiquetación del Lenguaje Natural (robust syntactic analysis methods for natural language tagging)*. Doctoral thesis, Universidad de A Coruña, Spain.

MOLINERO M. A., BARCALA F. M., OTERO J. & GRAÑA J. (2007). Practical application of one-pass viterbi algorithm in tokenization and pos tagging. *Recent Advances in Natural Language Processing (RANLP). Proceedings, pp. 35-40*.

NICOLAS L., FARRÉ J. & VILLEMONTE DE LA CLERGERIE É. (2007a). Confondre le coupable. In *Proceedings of TALN'07*, p. 315–324, Toulouse, France.

NICOLAS L., FARRÉ J. & VILLEMONTE DE LA CLERGERIE É. (2007b). Correction mining in parsing results. In *Proceedings of LTC'07*, Poznan, Poland.

NICOLAS L., SAGOT B., MOLINERO M. A., FARRÉ J. & VILLEMONTE DE LA CLERGERIE E. (2008). Computer aided correction and extension of a syntactic wide-coverage lexicon. In *Proceedings of Coling 2008*, Manchester.

SAGOT B. (2005). Automatic acquisition of a Slovak lexicon from a raw corpus. In *Lecture Notes in Artificial Intelligence 3658 (© Springer-Verlag), Proceedings of TSD'05*, p. 156–163, Karlovy Vary, République Tchèque.

SAGOT B. & DE LA CLERGERIE E. (2008). Fouille d'erreurs sur des sorties d'analyseurs syntaxiques. *Traitement Automatique des Langues*, **49**(1). (to appear).

SAGOT B. & VILLEMONTE DE LA CLERGERIE É. (2006). Error mining in parsing results. In *Proceedings of ACL/COLING'06*, p. 329–336, Sydney, Australia.

THOMASSET F. & VILLEMONTE DE LA CLERGERIE É. (2005). Comment obtenir plus des méta-grammaires. In *Proceedings of TALN'05*.

VAN DE CRUYS T. (2006). Automatically extending the lexicon for parsing. In *Proceedings of the eleventh ESSLLI student session*.

VAN NOORD G. (2004). Error mining for wide-coverage grammar engineering. In *Proceedings of ACL 2004*, Barcelona, Spain.

YI Z. & KORDONI V. (2006). Automated deep lexical acquisition for robust open texts processing. In *Proceedings of LREC-2006*.